# Data Analysis and Visualization in R (IN2339)

Exercise Session 7 - Statistical Testing I

Felix Brechtmann, Jun Cheng, Vicente Yepez, Julien Gagneur

## Section 00 - Getting Ready

1. Make sure you have already installed and loaded the following libraries:

```r
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(patchwork) # optional, makes plots nicer
```

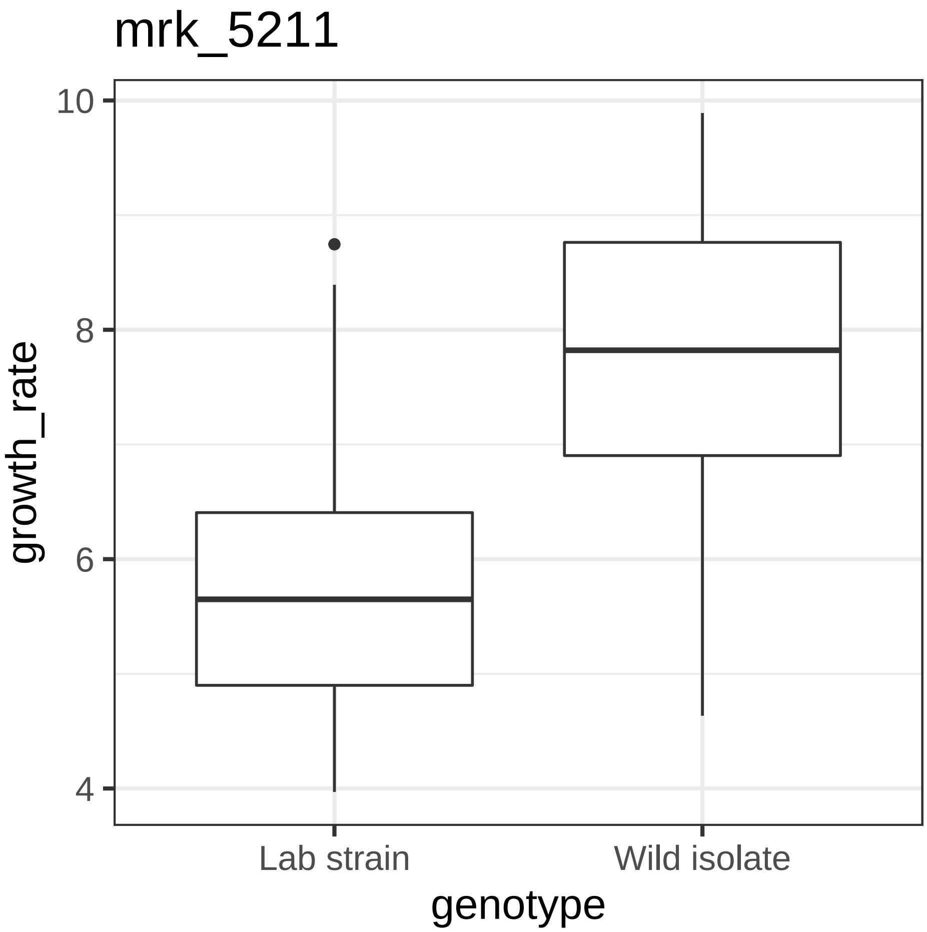
2. Load the yeast data

```r
genotype <- fread("./extdata/eqtl/genotype.txt")
genotype <- melt(genotype, id.vars = 'strain', variable.name = 'marker',
                 value.name = 'genotype')
growth <- fread("./extdata/eqtl/growth.txt")
growth <- melt(growth, id.vars = "strain", variable.name = 'media',
               value.name = 'growth_rate')
marker <- fread("./extdata/eqtl/marker.txt")
```

## Section 01 - Permutation test of growth rate difference

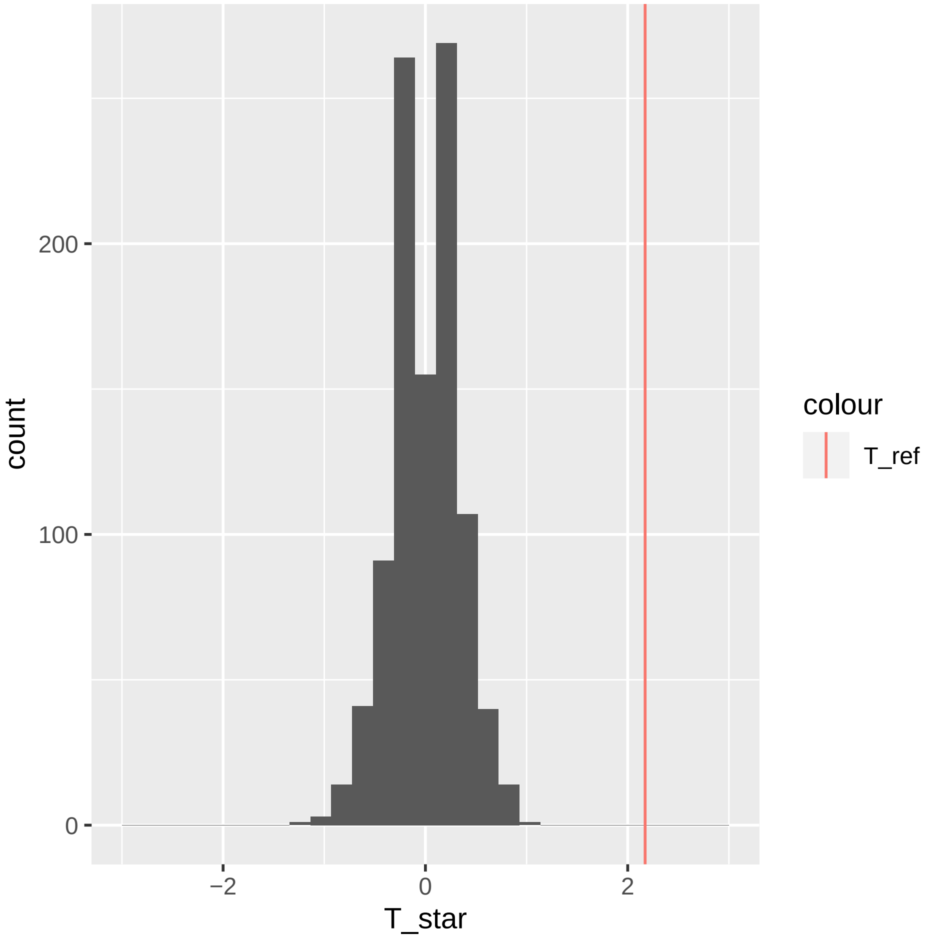
1. The following code recreates the example shown in the lecture to test the association of the genotype at marker 5211 with the growth rate difference in Maltose medium. Note that the code is written using functions, meaning that it will work for any marker, not just marker 5211. Read it carefully to understand what happens in each function. Then execute the code.

```r
# Plotting the growth rate difference
getMaltoseDt = function(mrk){
  growth_mrk <-  merge(growth, genotype[marker %in% mrk, .(strain, genotype, marker)],
                        by = 'strain', allow.cartesian = TRUE)
  growth_mrk[media == "YPMalt"]
}

# boxplot
plot_growth_one_mk <- function(mk){
    ggplot(getMaltoseDt(mk), aes(genotype, growth_rate)) +
    geom_boxplot() +
    labs(title = mk) + theme_bw(base_size = 16)
}
plot_growth_one_mk("mrk_5211")
```



```
# Function to calculate the difference of the median of two genotypes
median_diff <- function(dt){
  dt[genotype == ’Wild isolate’, median(growth_rate, na.rm=T)] -
    dt[genotype == ’Lab strain’, median(growth_rate, na.rm=T)]
}

# Function to permute the table, plot the resulting histogram
# and compute a p-value
p_val_medians <- function(dt, N_permu = 1000){
  # It will return both a pvalue and plot a histogram of T_star
  T_ref <- median_diff(dt)
  T_star <- sapply(1:N_permu, function(x){
      median_diff(dt[, genotype := sample(genotype)]) })
  # Plot
  g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) + geom_histogram() +
    geom_vline(aes(xintercept=T_ref, color="T_ref")) + xlim(-3,3)
  print(g) # Needed to render plot inside function call
  # Compute and return the p value
  p_val <- (sum(T_star > T_ref | T_star < -T_ref) + 1) / (N_permu + 1)
  p_val
}

# Calling the function:
```

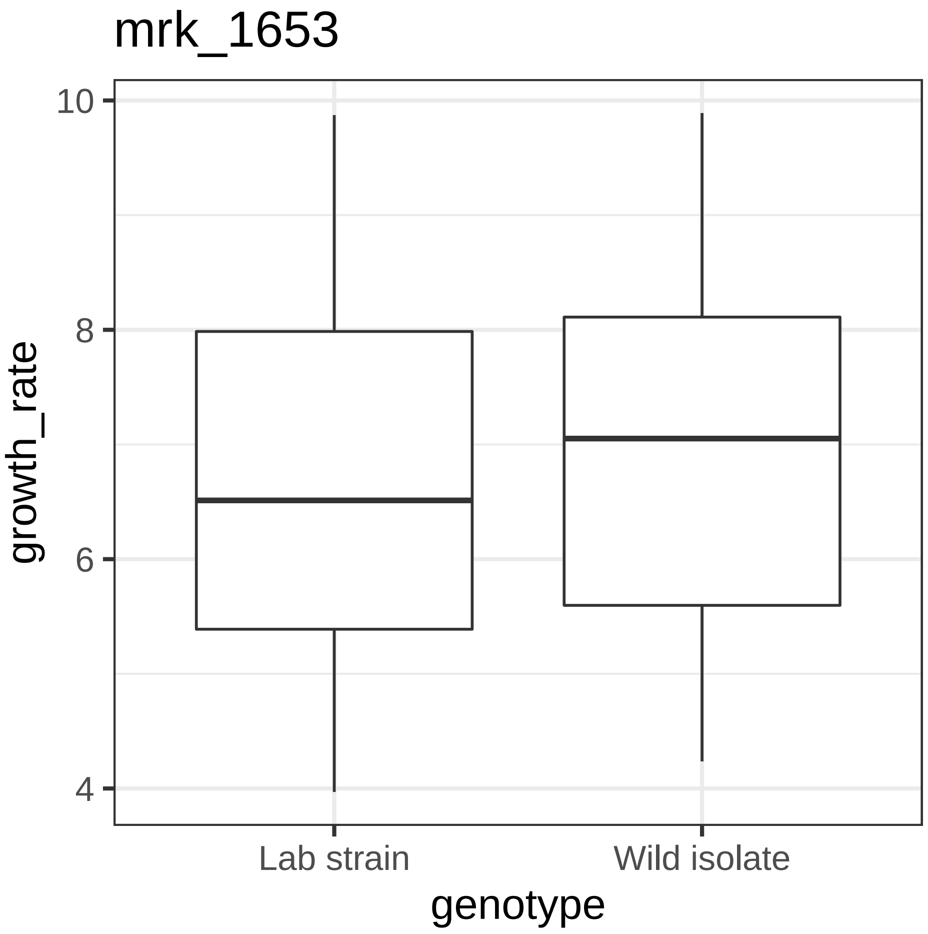
```
p_val_medians(getMaltoseDt("mrk_5211"))
```



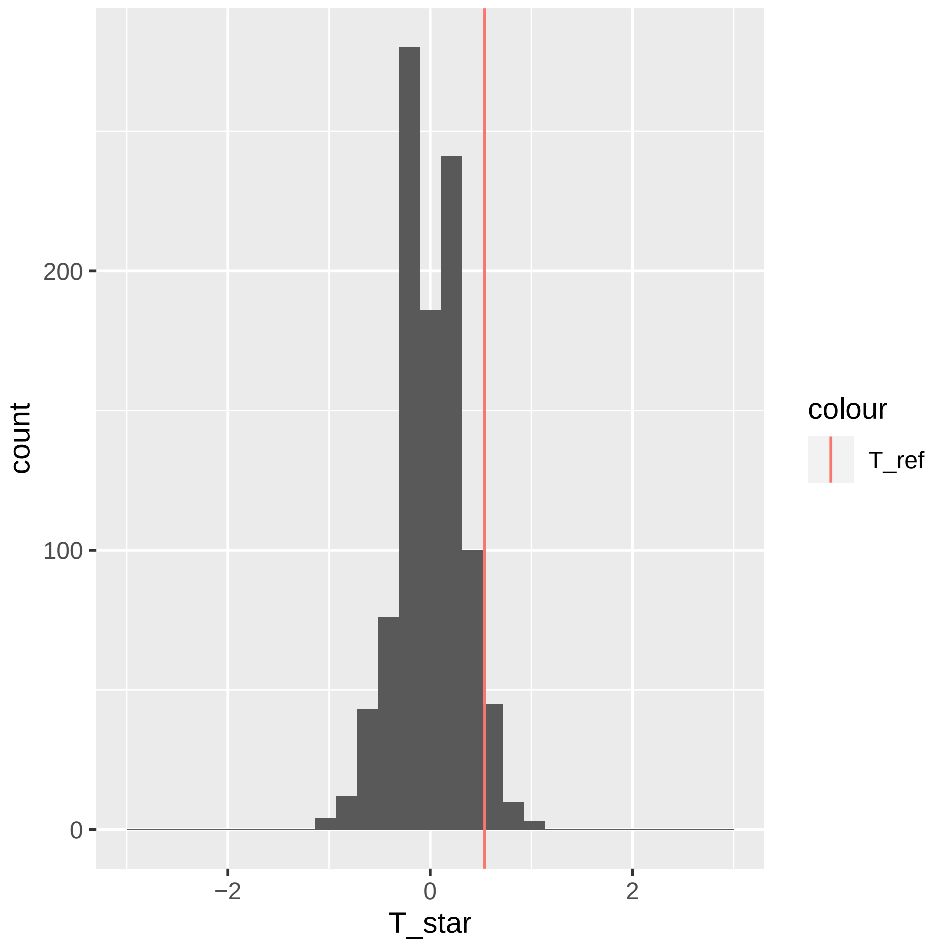
```
## [1] 0.000999001
```

2. Using the code above, plot and test whether markers 1653 and 5091 associate with growth. Interpret your results.

```
plot_growth_one_mk("mrk_1653")
```

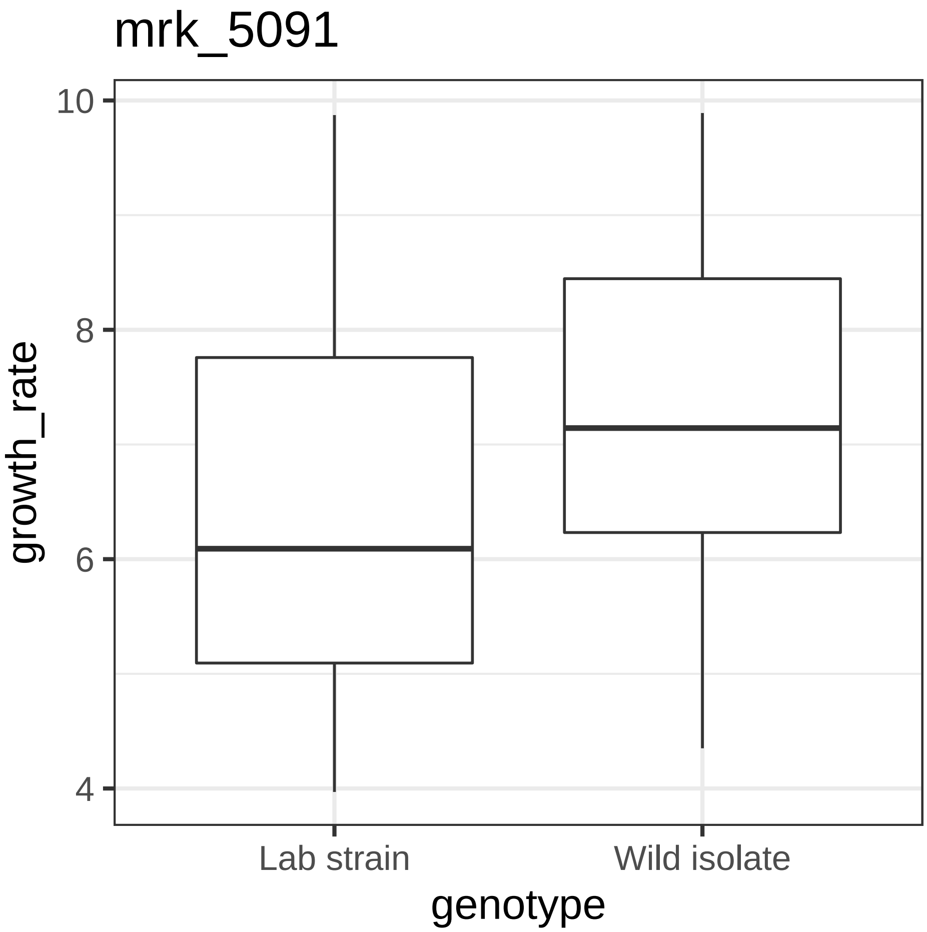


```
p_val_medians(getMaltoseDt("mrk_1653"))
```

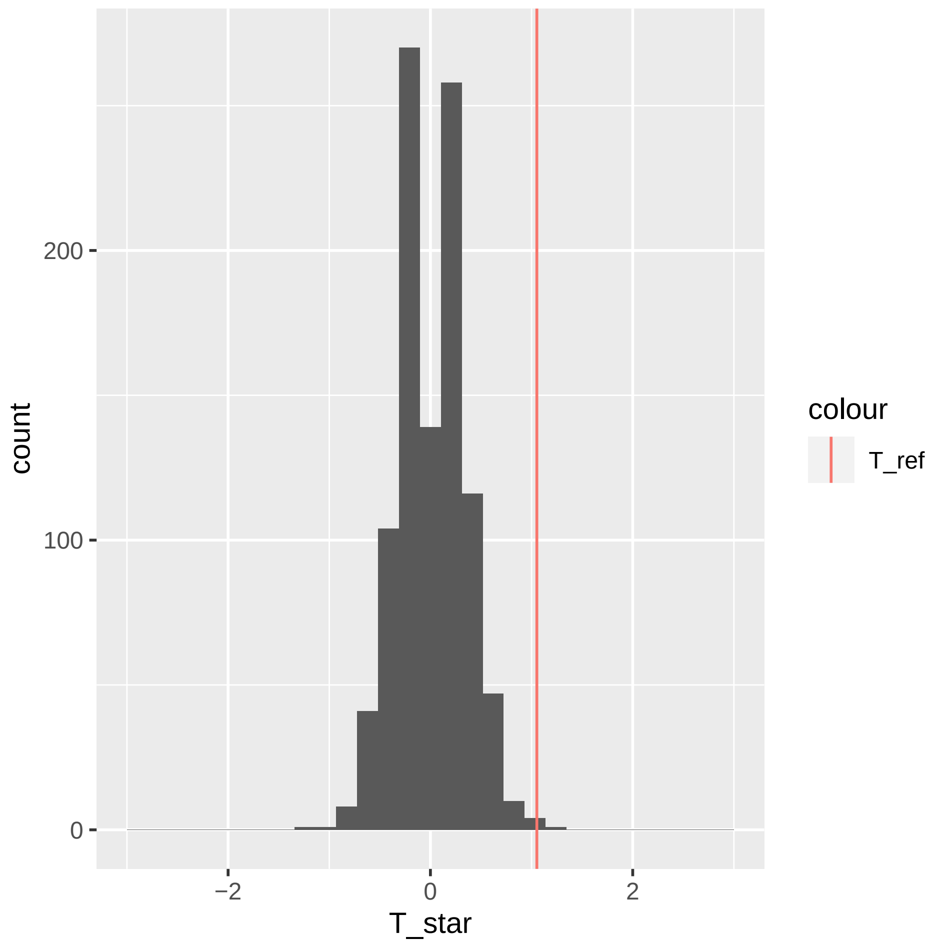


```
## [1] 0.0999001
```

```r
plot_growth_one_mk("mrk_5091")
```



```
p_val_medians(getMaltoseDt("mrk_5091"))
```



```
## [1] 0.004995005
```

```
# Marker 1653 is not significantly associated with growth, but marker 5091 is.
```

## Section 02 - Permutation test of marker association

1. We just concluded that both markers 5211 and 5091 are significantly associated with growth. However, this could be confounded. A common source of confounding in genomics is due to "linkage", which describes the phenomenon of markers being inherited together.

To investigate the issue of linkage in our dataset, test if marker 5091 significantly associates with marker 5211. Define a null hypothesis, a statistics and use permutation testing to answer the question. Strengthen your answer with a relevant plot.

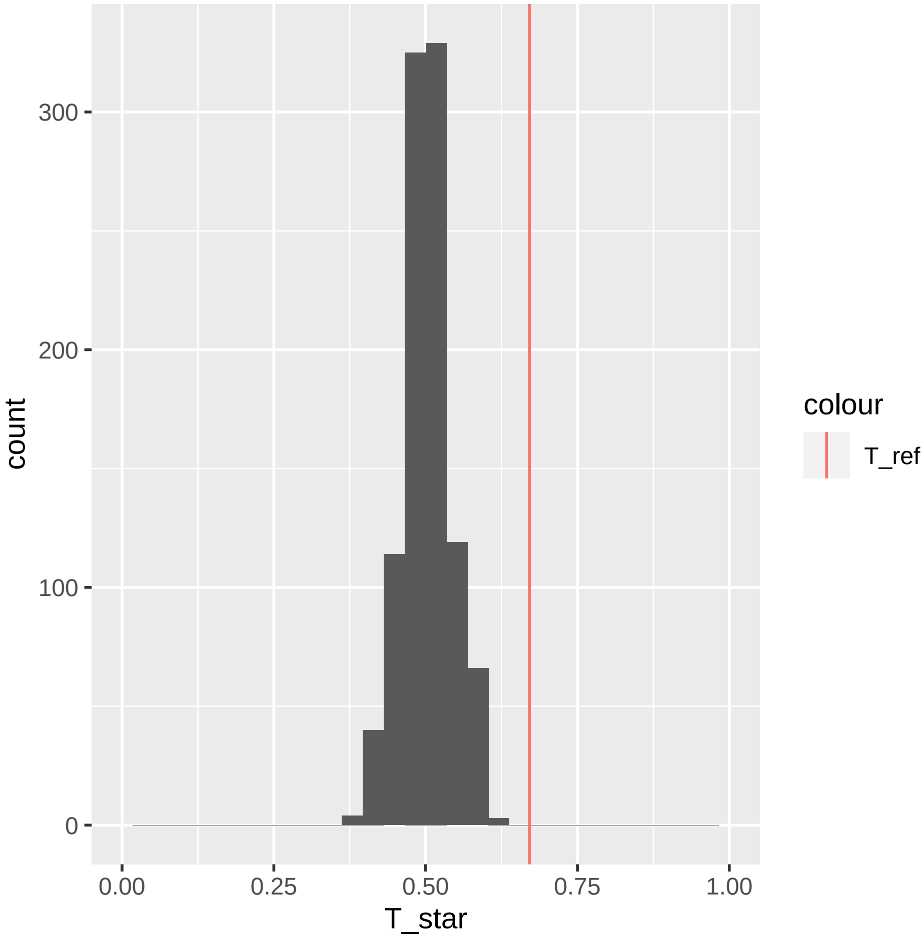
**Hint:** start from:

```
mks_geno <- genotype[marker %in% c('mrk_5091', 'mrk_5211')] %>%
  spread(marker, genotype)
```

and think about how this can be permuted.

```
# Ho: Marker 5091 is not significantly associated with marker 5211
# T statistic: number of times both markers had the same genotype
# OR
# T statistic: number of times both markers had the same genotype / number of strains
```

```
mks_geno <- genotype[marker %in% c(’mrk_5091’, ’mrk_5211’)] %>%
  spread(marker, genotype)

# Compute the number of times both markers had the same genotype
#T_ref <- mks_geno[mrk_5091 == mrk_5211, .N] # First option
T_ref <- mks_geno[mrk_5091 == mrk_5211, .N]/nrow(mks_geno) # Second option

# permutation
N_permu <- 1000
T_star <- sapply(1:N_permu, function(x){
    mks_geno[mrk_5091 == sample(mrk_5211), .N]/nrow(mks_geno)})   # other alternative

# plot distribution
ggplot(data = data.table(T_star = T_star), aes(T_star)) + geom_histogram() +
  geom_vline(aes(xintercept=T_ref, color="T_ref")) + xlim(0,1)
```



```
# compute p-value
p_val <- (sum(T_star > T_ref) + 1) / (N_permu + 1)
p_val
```

```
## [1] 0.000999001
```

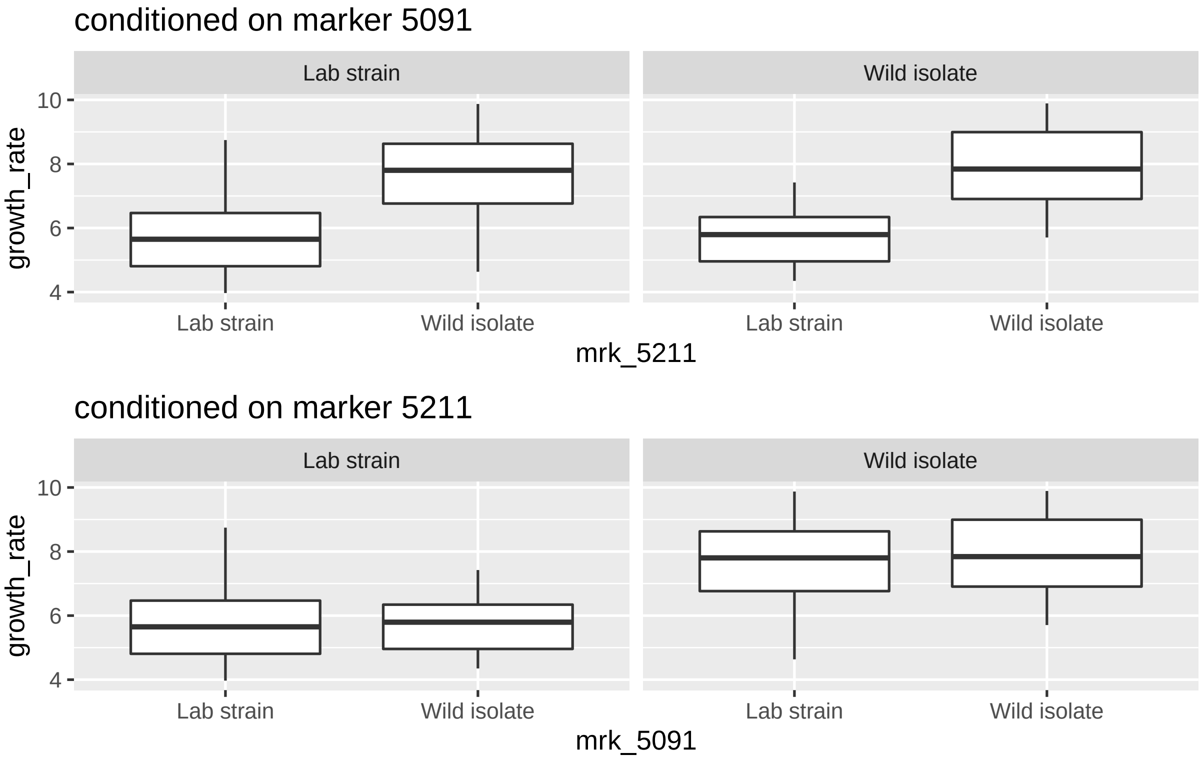
```
# The two markers are significantly associated
# It appears that they are often inherited together
```

## Section 03 - Accounting for Confounding

1. We see that indeed marker 5211 and 5091 associate. Thus, the assocation between these markers and growth could be confounded.

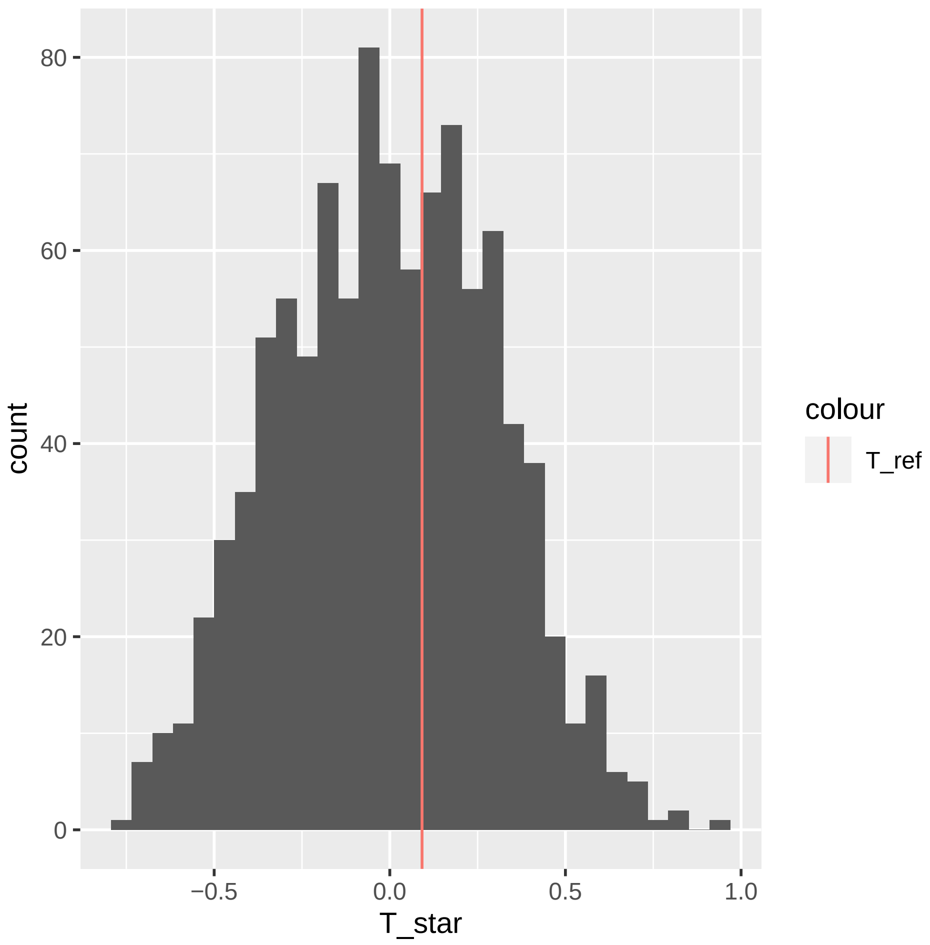
We now would like to know if marker 5091 still associates with growth in maltose (YPMalt) when conditioned on marker 5211. Define a null hypothesis, a statistics and use permutation testing to answer the question. Strengthen your answer with a relevant plot.

```
# Add growth in maltose (YPMalt) to the genotype data
conditioning_dt <- merge(mks_geno, growth[media == 'YPMalt'], by = 'strain')

a <- ggplot(conditioning_dt, aes(mrk_5211, growth_rate)) +
    geom_boxplot() +
    facet_wrap(~ mrk_5091) +
    labs(title='conditioned on marker 5091')

b <- ggplot(conditioning_dt, aes(mrk_5091, growth_rate)) +
    geom_boxplot() +
    facet_wrap(~ mrk_5211) +
    labs(title='conditioned on marker 5211')
a / b #Patchwork syntax to nicely align plots
```



```
# Boxplots give a hint about the confounding
# We see that the effect of mrk 5211 persists
# even when we condition on the other marker
# This is not true for mrk 5091
# But we should test!
```

```
p_val_condition_on <- function(test_mrk = "mrk_5078", condition_mrk = "mrk_5211", N_permu = 1000) {
  # On the simple growth vs genotype case:
  ## Ho: For each marker, the growth medians are the same for Lab and Wild
  ## Tref: median(growth on Wild) - median(growth on Lab), for each marker

  # On the growth vs genotype case, conditioned on another marker:
  ## Ho: For each marker, the growth medians are the same for Lab and Wild,
  ## no matter the conditioned marker
  ## Tref: mean across subgroups of {median(growth on Wild) - median(growth on Lab)},
  ## for each marker

  # Prepare data table
  conditioned_dt <- getMaltoseDt(c(test_mrk, condition_mrk)) %>%
    spread(marker, genotype)
  setnames(conditioned_dt, test_mrk, "test_mrk")
  setnames(conditioned_dt, condition_mrk, "condition_mrk")
  # Get T_ref
  median_ref <- conditioned_dt[, median(growth_rate, na.rm=T), by = c("test_mrk", "condition_mrk")] %>%
        spread(test_mrk, V1)
  T_ref <- mean(median_ref[, ‘Wild isolate‘ - ‘Lab strain‘])
  # Do permutations conditioned on the other marker
  T_star <- numeric(N_permu)
  for(i in 1:N_permu){
    conditioned_dt[, test_mrk := sample(test_mrk), by = condition_mrk]
    medians <- conditioned_dt[, median(growth_rate, na.rm=T), by = c("test_mrk", "condition_mrk")] %>%
        spread(test_mrk, V1)
    T_star[i] <- mean(medians[, ‘Wild isolate‘ - ‘Lab strain‘])
  }
  # Plot
  g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) + geom_histogram() +
      geom_vline(aes(xintercept=T_ref, color="T_ref"))
  print(g)
  # P-value
  p_val <- (sum(T_star > T_ref) + 1) / (N_permu + 1)
  p_val
}
p_val_condition_on(test_mrk = "mrk_5091", condition_mrk = "mrk_5211")
```

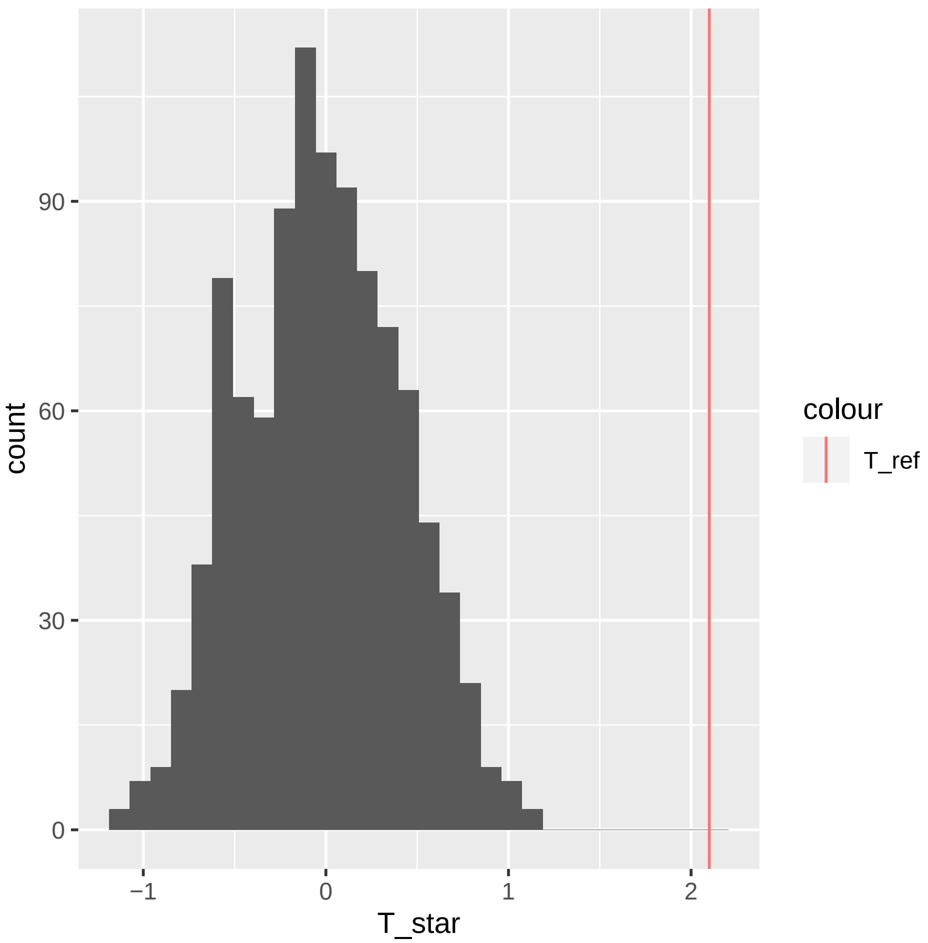


```
## [1] 0.3946054
```

```
# No significant effect of mrk 5091 on growth once we condition on 5211
```

2. Now, test if marker 5211 associates with growth in maltose when conditioned on marker 5091. Are the results the same? Discuss.

```
p_val_condition_on(test_mrk = "mrk_5211", condition_mrk = "mrk_5091")
```



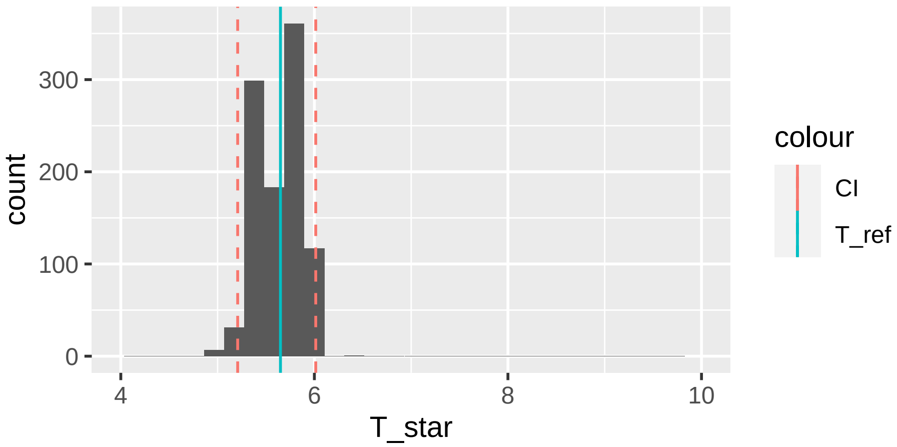
```
## [1] 0.000999001
```

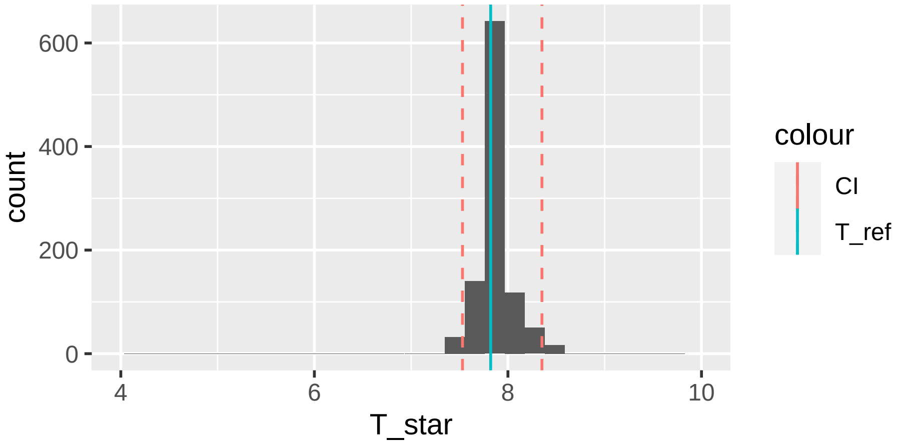
```
# 5211 still has an effect, even when conditioning on 5091
# This provides some evidence that 5211 is the "causal" marker
# But we have not excluded every possible source of confounding
# So we should not generalize too much
```

### Section 04 - Confidence Intervals

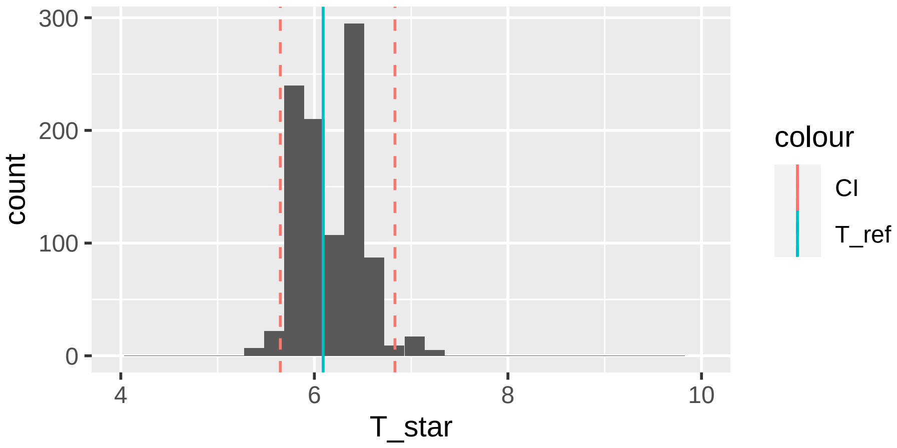
1. Estimate 95% equi-tailed confidence intervals for the median of growth in maltose for each genotype at marker mrk_5211. Use the case resampling bootstrap scheme and report bootstrap percentile intervals. Propose a visualization of the results. Try it also with markers 5091 and 1653.

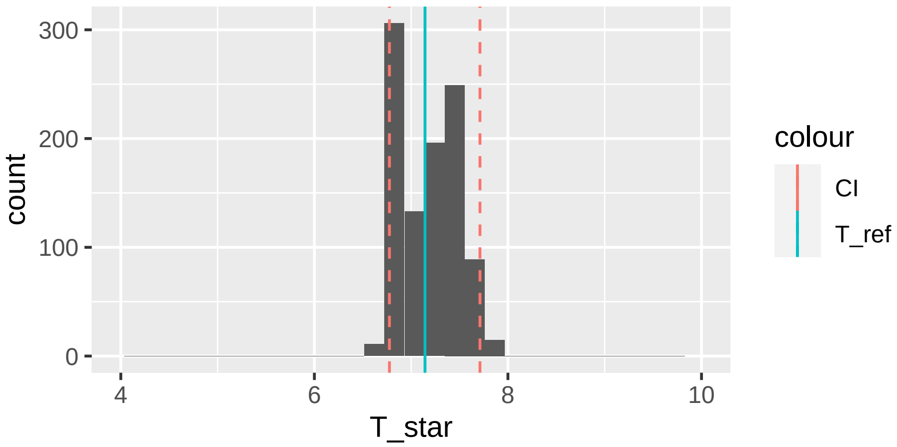
```
mystat <- function(x){
  median(x, na.rm=TRUE)
}

# Bootstrap and compute some function func
boot <- function(x, func, B = 999){
  T_star <- sapply(1:B, function(i){
      xstar <- sample(x, replace=TRUE)
      func(xstar)
      }
  )
  return(T_star)
}

confint <- function(Tstar, alpha = 0.05){
  quantile(Tstar, c(alpha/2, 1-alpha/2))
}

conf_int_plot <- function(marker){

  plot_list <- list("Lab strain" = NA, "Wild isolate" = NA)
  for(geno in c("Lab strain", "Wild isolate")){
    # geno = 'Lab strain'
    x <- getMaltoseDt(marker)[genotype == geno, growth_rate]
    T_star <- boot(x , mystat)  # Bootstrap 1000 times and compute the median (mystat)
    T_ref <- median(x, na.rm=TRUE)
    CI_lab <- confint(T_star)
    # Plot histogram, add median and confidence interval as vertical lines
    g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) + geom_histogram() +
      geom_vline(data=data.table(T_ref), aes(xintercept=T_ref, color="T_ref")) + xlim(4,10) +
      geom_vline(data=data.table(CI_lab), aes(xintercept=CI_lab[1], color="CI"), linetype="dashed") +
      geom_vline(data=data.table(CI_lab), aes(xintercept=CI_lab[2], color="CI"), linetype="dashed")
    plot_list[geno] <- list(g) # list is necessary to let patchwork interpret it right
  }
  # this is patchwork syntax
  # it nicely aligns plots above each other
  # this is completely optional
  plot_list[["Lab strain"]] / plot_list[["Wild isolate"]]
}

conf_int_plot("mrk_5211")
```



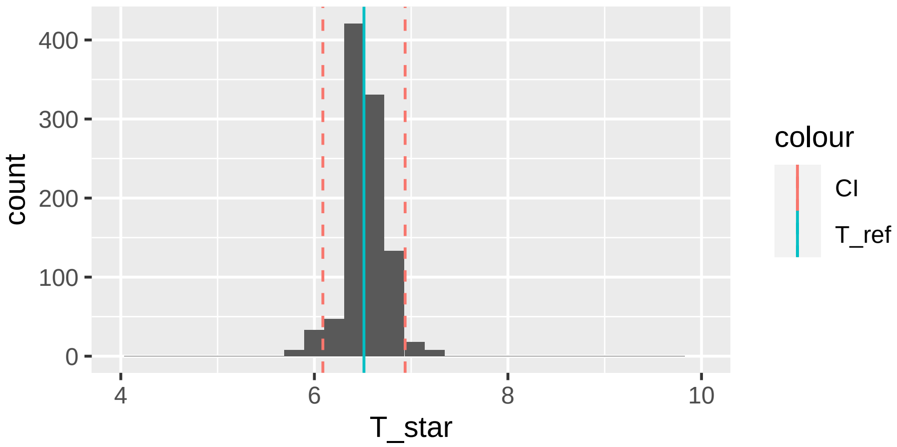


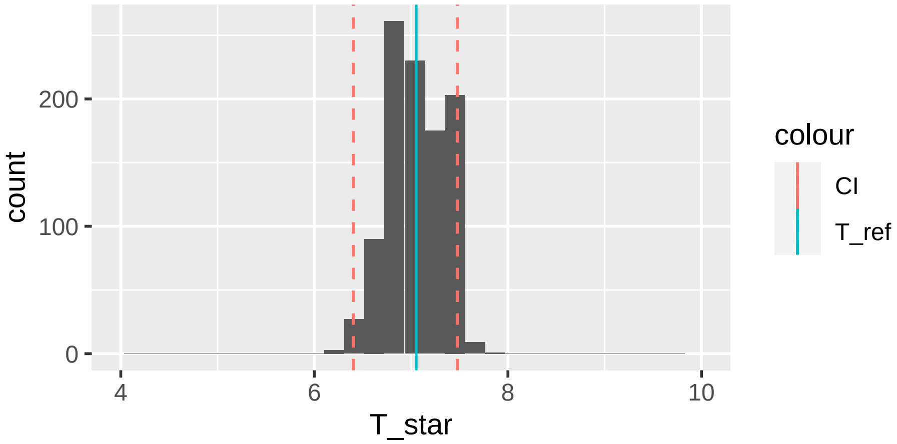
```r
# We see that the confidence intervals are very far
conf_int_plot("mrk_5091")
```





```r
# The intervals are quite close
conf_int_plot("mrk_1653")
```





```
# The intervals clearly overlap

# Note: saying that the confidence intervals for the median growth rates overlap
# Does *NOT* necessarily mean that we do not reject the null hypothesis of
# the tests conducted above for the difference in median growth rates.
# This is because the confidence interval for the difference in median growth rates
# may not contain zero, even if the individual CI for the median growth rates overlap.
# You can see this by noting, for two independent random variables X and Y
# that std(X-Y) = sqrt(Var(X-Y) = sqrt(Var(X) + Var(Y))
# And this will generally be smaller than sqrt(Var(X)) + sqrt(Var(Y))
# https://www.cmaj.ca/content/166/1/65.long
```